引用格式：刘鹏华,姚尧,梁昊,等. 耦合卡尔曼滤波和多层次聚类的中国PM 2.5时空分布分析[J].地球信息科学学报,2017,19(4):475-485. [Liu P H, Yao Y, Liang H, et al. 2017. Analyzing spatiotemporal distribution of $PM_{2.5}$ in China by integrating Kalman filter and multi-level clustering. Journal of Geo-information Science, 19(4):475-485.] **DOI:** 10.3724/SP.J.1047.2017.00475

耦合卡尔曼滤波和多层次聚类的中国$PM_{2.5}$时空分布分析

刘鹏华[1],姚　尧[1,2*],梁　昊[3],梁兆堂[1],张亚涛[1],王昊松[1]

1. 中山大学地理科学与规划学院,广州 510275; 2. 中山大学 广东省城市化与地理环境空间模拟重点实验室,广州 510275;
3. 南京大学 江苏省地理信息技术重点实验室,南京 210023

Analyzing Spatiotemporal Distribution of $PM_{2.5}$ in China by Integrating Kalman Filter and Multi-level Clustering

LIU Penghua[1], YAO Yao[1,2*], LIANG Hao[3], LIANG Zhaotang[1], ZHANG Yatao[1] and WANG Haosong[1]

1. School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China; 2. Key Laboratory for Urbanization and Geo-simulation of Guangdong Province, Sun Yat-sen University, Guangzhou 510275, China; 3. Key Laboratory for Geographical Information Science and Technology of Jiangsu Province, Nanjing University, Nanjing 210023, China

Abstract: Serious air pollution has recently aroused wide public concerns in China. The traditional method of quantitative remote sensing model is not only sophisticated but also inaccurate to fetch the exact $PM_{2.5}$ data near the ground. Though the built-up ground monitoring stations can now provide sufficient $PM_{2.5}$ observation data with high sampling frequency, there still exist many extreme outliers due to inevitable observation noise. Therefore, in this study, we adopted Kalman filter for optimal estimation of time-series of air quality data in 338 cities of China and comprehensively analyzed the spatiotemporal distribution pattern during the period of 2015. In our detailed analysis, we used DTW based K-Medoids clustering to classify cities into 4 levels according to their contamination degree, and utilized q statistic technique to evaluate the spatial stratified heterogeneity of $PM_{2.5}$. The results show that by using Kalman filter, noise can be effectively reduced and value of PSNR can be significantly improved. In the study of temporal distribution, we found that $PM_{2.5}$ followed a 'U' curve in yearly temporal distributions while daily temporal distributions obeyed a 'W' curve. $PM_{2.5}$ density is much higher in winter than in summer in China, and spatial stratified heterogeneity is even more pronounced during the fall-winter stage. In the study of spatial distribution, it can be clearly seen that $PM_{2.5}$ appears a 'Dual-core' pattern across China where concentration of $PM_{2.5}$ spiked at Xinjiang and North China plain. In contrast, Xizang, Guangdong and Yunnan are more stable areas with excellent air quality, ranking first-tier nationwide.

Key words: $PM_{2.5}$; big Data; kalman filter; spatiotemporal analysis; K-Medoids

***Corresponding author:** YAO Yao, E-mail: whuyao@foxmail.com

摘要：近年来,细颗粒物污染尤其是$PM_{2.5}$受到人们越来越多的关注,研究$PM_{2.5}$的时空分布规律也具有越来越重大的意义。传

收稿日期：2016-07-01;**修回日期：**2016-11-01.
基金项目：国家自然科学重点基金项目(41531176);国家自然科学基金项目(41671398、41601420)。
作者简介：刘鹏华(1995-),男,本科生,研究方向为遥感与地理信息系统。E-mail: liuphhhh@foxmail.com
***通讯作者：**姚　尧(1987-),男,博士生,研究方向为时空大数据分析和精细城市模拟。E-mail: whuyao@foxmail.com

统的遥感反演方法模型复杂,且不能揭示近地表面的$PM_{2.5}$分布规律。地面监测站的建设为$PM_{2.5}$的研究提供了更实时的观测数据,但由于测量噪声的影响,观测数据存在不准确的极端异常值。为了揭示中国$PM_{2.5}$的时空分布特征,本研究采用Kalman滤波对2015年中国338个城市的空气质量监测网络大数据进行最佳估计,并分析其时空特征。同时,根据中国各城市的$PM_{2.5}$浓度的时序分布,采用基于DTW的K-Medoids聚类方法将其分为4个等级,并采用q统计量来评估$PM_{2.5}$浓度分布的空间分层异质性。结果表明,采用Kalman滤波能有效去除数据噪声,峰值信噪比(PSNR)明显增大。在时空分布上,中国$PM_{2.5}$时间分布曲线呈现"U"形,冬季$PM_{2.5}$浓度明显高于夏季,且日变化曲线呈现"W"形;秋冬季$PM_{2.5}$浓度的空间分层异质性非常显著,且空间分布呈现"双核分布",重污染区主要分布在华北平原、新疆等地,西藏、广东、云南等地是稳定的空气质量优良区。

关键词:$PM_{2.5}$;大数据;卡尔曼滤波;时空分析;K-Medoids

1 引言

近年来,中国频繁发生连续高强度的雾霾天气和大气污染,$PM_{2.5}$已经成为主要的大气污染物[1]。$PM_{2.5}$能对可见光产生消光作用,从而使可见度降低,更严重的是它进入人体后沉积在肺泡和支气管,对人体健康造成危害[2-4]。为此,研究$PM_{2.5}$的分布特征具有重大意义。

传统的$PM_{2.5}$的研究主要利用卫星观测数据,通过遥感反演气溶胶光学厚度(AOT)来揭示$PM_{2.5}$的分布规律[5-8]。但是,利用遥感反演方法数据更新周期长,难以揭示不同时间尺度(季节、月、日)的$PM_{2.5}$浓度变化规律,也难以反映近地面$PM_{2.5}$浓度的空间分布格局[9]。2012年以来,中国陆续在全国各城市建设了空气质量监测站,并实时监测和发布$PM_{2.5}$等6项污染物浓度数据。监测站的数据更新周期为1 h,因而能揭示不同时间尺度的$PM_{2.5}$变化规律。武装[10]等利用监测数据,基于Hadoop平台进行了空气污染时空分布的可视化分析,王振波等[11]基于中国2014年190个城市的945个监测站的$PM_{2.5}$浓度观测数据,采用空间数据统计模型,揭示了中国$PM_{2.5}$的时空分布格局。然而,时空维度上的直观统计量只能揭示特定城市的时间变化规律,而不能反映其与其他城市的相关性。而且,由于测量噪声的存在,采用均值滤波处理监测站数据不能反映$PM_{2.5}$浓度的真实分布。

为了真实反映$PM_{2.5}$的时空分布特征和城市之间的相关性,本研究采用网络开源大数据,基于一维线性卡尔曼滤波获取$PM_{2.5}$浓度的最佳估计值,分析和揭示中国$PM_{2.5}$的时空分布规律;并根据城市$PM_{2.5}$浓度的月度变化规律采用基于DTW的K-Medoids方法将中国城市聚为4类,通过研究每类城市的特征,进一步揭示了不同城市之间的$PM_{2.5}$分布的相似性和关联性。

2 研究区概况与数据源

2.1 研究区概况

本文的研究区为中国大陆,包括23个省、5个自治区、4个直辖市、2个特别行政区。近年来,中国频繁遭遇严重的雾霾天气和连续高强度的大气污染[12]。随着《环境空气质量标准》的出台,中国陆续建设了一批空气质量监测站,并实时向公众提供空气质量信息。如图1所示,截止2015年12月,研究区内1406个空气质量监测站点分布于全国338个城市。监测站点集中分布在社会经济较发达的京津冀地区和东部沿海省份,其中,京津、山东、江苏、广东的监测站点分布最密集,约占34.47%,而新疆、西藏、青海等西部省份虽然地域宽广,监测站点却极稀疏,仅占5.08%。

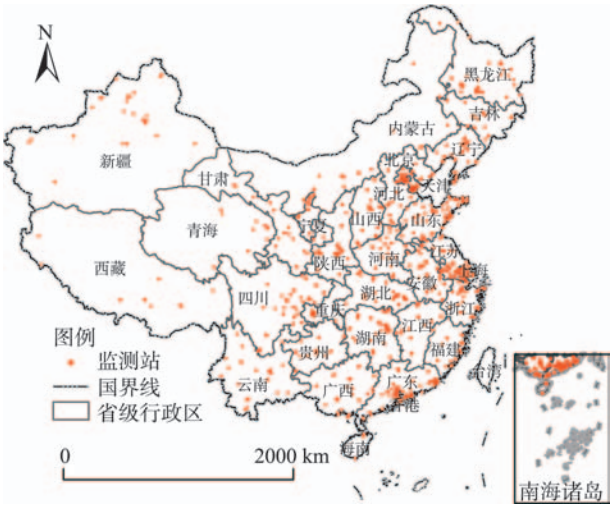

图1 中国空气质量监测站点分布图

Fig.1 Air quality monitoring stations of China

2.2 数据源

本文研究数据来自BestApp工作室从环保部获取监测数据而建立的空气质量数据共享平台(http://pm25.in/),该平台逐小时发布空气质量监测数据,并提供免费的$PM_{2.5}$数据接口。本文研究数据均从

该平台数据接口获取,包括2015年1−12月的1200多万条数据,每条数据包括空气质量指标(AQI),CO、NO_2、O_3、SO_2、颗粒物等浓度,以及站点所在城市、数据发布时间等说明性记录。

3　网络$PM_{2.5}$监测数据的时空分析

3.1　基于Kalman滤波的$PM_{2.5}$数据清洗

由于各个站点投入使用的时间不一致,且监测环境不同,数据存在大量的缺失。一方面,新的监测站点在不断建设和工作,导致不同时间所获取的空气质量监测数据站点数目不一致,新建站点在发布数据之前的数据存在缺失。另一方面,由于不同站点的监测环境差异,部分站点在某些时刻未能及时发布新数据,或发布的数据存在字段缺失。据统计,以2015年12月31日的监测站点记录为标准,研究数据中存在约4.23%的数据缺失和0.76%的零值记录。为此,本研究对缺失数据进行线性内插处理。

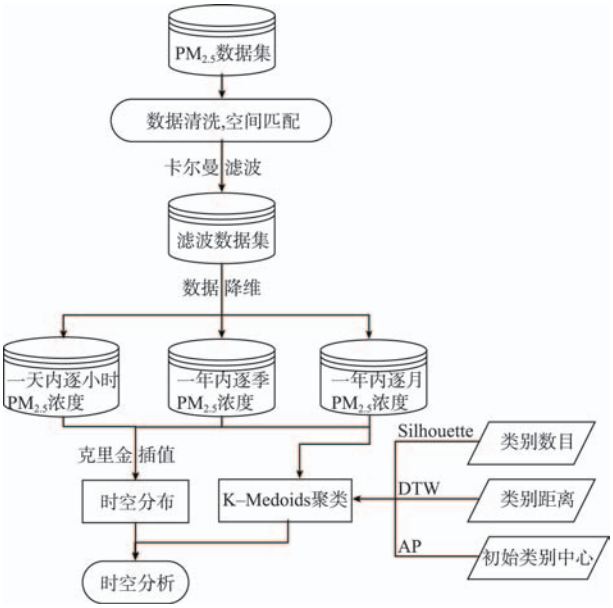

图2　研究流程图

Fig.2　Research workflow

在对缺失数据进行线性内插后,数据能保持较高的一致性。然而,在实际测量中,由于气候、自然环境、污染源的差异[13],监测站采用的微量振荡天平法和Beta射线法都存在监测数据偏差太大和不稳定的问题[14]。为此,本研究采用Kalman滤波对观测进行最佳估计,进而对时序数据进行降维处理(图2)。Kalman滤波是Kalman[15−16]提出的一种时域滤波算法,其采用时间递推的方式,考虑了系统的过程噪声和测量噪声[17],是一种对观测值的线性最小方差估计方法[15]。Kalman滤波可以基于系统上一时刻的状态预测下一状态,当获得下一状态的观测值时,根据下一状态的预测结果和观测结果获得下一状态的最优化估计。由于在状态预测和最优估计更新时状态的噪声也被更新,因此Kalman不仅能够处理平稳变化的随机过程,也能处理多维和非平稳的随机过程[18]。

图3给出了基于Kalman滤波对$PM_{2.5}$时序观测数据进行滤波处理的伪代码。其中,系统参数A为状态转移矩阵,表示后一状态对前一状态的影响;B为控制输入矩阵,描述后一状态驱动因素对后一状态的影响;U为驱动输入向量,即与$PM_{2.5}$浓度分布

```
Algorithm KalmanFiltering(A[n])
  Input:
    A[n]为n个观测数据，即一组PM2.5监测站观测值
Initialize:
  初始化系统过程噪声W和测量噪声V，系统参数A、B和控制量U;
  设置系统的初始最优值B[1]为第一个观测值A[1];
for (i=2; i<=n; i++)
  根据 i-1 状态的观测值 A[i-1] 和最优值 B[i-1] 以及控制参数 A、B、U 来计算 i 状态的预测值 X[i];
  基于 i-1 状态最优值的误差 PB[i-1] 和系统过程噪声的误差 Q 来更新 i 状态预测值的误差 PX[i];
  根据测量噪声的误差 R 和 PX[i] 来计算 i 状态的卡尔曼增益 Kg[i];
  收集 i 状态的观测值后，结合卡尔曼增益 Kg[i] 、预测值 X[i] 和测量值 A[i] 来计算 i 状态下的最优
  值B[i];
  更新 i 状态下最优值的误差 PB[i];
End for
Calculate PSNR:
  计算 B 的均值 b;
  以b 为真实值，计算 MSE;
  计算 PSNR;
Output:
  经过滤波后 A 的最优估计值 B 以及 PSNR。
```

图3　Kalman滤波伪代码

Fig.3　Pseudo code of Kalman filter

相关的因子对其影响；Kg表示卡尔曼增益，它的值越大，表明真实值（即最优值）越接近测量值，反之越接近预测值。按照图3的处理流程，本研究对中国338个城市2015年1−12月的数据进行Kalman滤波，获得了对原始观测数据的最优估计，进而获得每小时、每个月、每个季度、全年平均$PM_{2.5}$浓度。另外，基于Kalman滤波计算出滤波前后数据集的均值，并将均值分别作为数据的真实参考值，用峰值信号噪声比（PSNR）来评估采用Kalman滤波结果作为数据集代表的有效性。

3.2 K-Medoids时序聚类

在Kalman滤波的基础上，本研究获取了中国各城市$PM_{2.5}$浓度的月度变化。为了有效分析不同地区$PM_{2.5}$浓度的时间分布差异，本研究采用基于动态时间规整（Dynamic Time Warping, DTW）的K-Medoids聚类方法，对各个城市的$PM_{2.5}$浓度分布分级划分。

K-Medoids聚类是Kaufman[19−20]提出的一种用数据相似度中心来表示聚类中心的聚类方法。相比于K-Means的最小化数据点之间欧氏距离的目标，K-Medoids的目标是使数据特征之间的相似度最小化，因而它对于噪声和异常值具有较强的稳健性。

为了描述不同城市$PM_{2.5}$浓度时间序列之间的相似度，常用的方法是通过离散傅里叶变换和离散小波变换进行降维变换[21−22]，并采用欧氏距离来评估其相似度，这些方法虽然可以获取时间序列之间的欧氏距离下界，然而，欧氏距离并不总能对数据之间的相似度进行很好的划分[23−25]，如比较$PM_{2.5}$浓度变化的时间序列时，需要特别关注浓度曲线的拐点，以浓度变化的峰值点、谷值点及其关系作为衡量序列之间相似性的重要根据。事实上，DTW可以衡量序列之间灵活的相似性和差异性，常被用于离散时间序列之间相似度的度量，并可在聚类算法中识别发现不同时间序列中的模式[26−28]。

同时，大气污染物的扩散受到地形复杂度、气象条件、大气污染物理化特征、污染源特征等多种因素的影响，其扩散过程十分复杂[29]，但通过计算任意2个区域$PM_{2.5}$浓度时间序列的DTW距离，可以衡量上述2个区域的$PM_{2.5}$变化模式的相似性。例如，由于大气污染物在从污染源向周围进行扩散时，随着时间推进，距离污染源较远位置的$PM_{2.5}$浓度峰值在时间轴上迟于距离污染源较近位置的$PM_{2.5}$浓度峰值，但二者的空气污染程度相似。而DTW可以发掘时间序列中的知识，寻找模式[27]，在引入聚类算法之后，具有相似知识模式的区域聚集在相同的簇中[28,30]，从而实现研究区域的$PM_{2.5}$聚类。

假设2个时间序列P和Q，$P=\{p_1,p_2,\cdots,p_i,\cdots,p_n\}$，$Q=\{q_1,q_2,\cdots,q_j,\cdots,q_m\}$，构建$P$和$Q$之间的相似度矩阵$M$，其中$M(i,j)=\left(p_i-q_j\right)^2(1\leqslant i\leqslant n, 1\leqslant j\leqslant m)$。DTW路径为$P$和$Q$之间的最佳映射，它是$M$矩阵中相邻元素的集合[31]，即$DTW=\{w_1,w_2,\cdots,w_k\}$，$w_t\in M(1\leqslant t\leqslant k)$，且$DTW$满足以下条件：

（1）$\max(m,n)\leqslant k\leqslant m+n-1$；

（2）$w_1=M(1,1)$，$w_k=M(n,m)$；

（3）如果w_t表示矩阵中的$M(i,j)$，那么w_{t+1}只能为$M(i+1,j)$或$M(i,j+1)$；

（4）DTW的路径长度d_{DTW}为所有满足上述条件的路径中的最短的，即$d_{DTW}=\min\left(\frac{\sum_{i=1}^{k}w_i}{k}\right)$。

本研究采用DTW对2个时间序列进行动态规整，并基于DTW路径的长度来衡量2个时间序列之间的相似度。同时，采用AP（Affinity Propagation）[32]算法初始化聚类中心，来减小随机选择对聚类结果的影响。AP算法在初始过程将所有数据点作为潜在的聚类中心，数据点之间通过吸引度和归属度之间的信息传递来竞争聚类中心和选择归属的聚类中心[33−34]，以此获得几个具有代表性的聚类中心。另外，为了精确化聚类数目，本研究引入Silhouette来评估聚类结果的合理性。Silhouette是Rousseeuw[35]提出的评价每个数据对象与其所属类别的适宜度的指标，它基于数据点与类内其他数据点和其他类中数据点之间的距离来衡量类内凝聚度和类间离散度[36]，从而确定最适宜的聚类数目[37]。对于Y类中的某个数据对象y_i，其Silhouette指标为：

$$S(y_i)\frac{b(y_i)-a(y_i)}{\max\{b(y_i),a(y_i)\}} \quad (1)$$

式中：$a(y_i)$为y_i与Y类中其他数据点之间的平均相似度；$b(y_i)$为其他类内所有数据点与y_i之间相似度的最小值。很显然，$s(y_i)$越高，表明数据y_i与Y类的适宜度越好。当所有类别中数据对象的平均Silhouette值最小时，聚类数目最适宜。

4 结果与讨论

4.1 $PM_{2.5}$时空分布分析

$PM_{2.5}$的观测值存在时间差异性，而且由于不同

城市监测环境和设备条件、自然社会环境、$PM_{2.5}$浓度分布不同,城市观测值的空间分布具有较大差异性。本研究基于Kalman滤波进行观测数据的最佳估计,为了评估Kalman滤波最佳估计与观测值的优劣,本研究选取北京、上海、广州、南京的24 h的$PM_{2.5}$浓度分析发现(表1),滤波后数据PSNR明显提高,信号失真度明显降低。因此,采用Kalman滤波对数据进行滤波处理能有效地去除噪声,更好地反映数据真实分布。

表1　Kalman滤波前后PSNR值对比表

Tab.1　PSNR value before and after Kalman filtering

地区	滤波前	滤波后
北京	12.8132	21.0545
上海	9.7523	14.9660
广州	13.2474	19.1636
南京	17.9053	27.0554

在采用Kalman滤波对数据进行最佳估计的基础上,本研究对缺失的数据进行线性内插,分别获取了中国各城市年度、季度、月度和逐小时的$PM_{2.5}$平均浓度,并采用Kriging插值法模拟中国12个月的$PM_{2.5}$空间分布,对中国的$PM_{2.5}$时空分布进行分析。

4.1.1　时间分布分析

(1)季度$PM_{2.5}$分布

2015年1−12月中国$PM_{2.5}$平均浓度为49 μg/m^3,按照中国环境空气质量(GB 3095-2012)$PM_{2.5}$浓度年、日均限值为15 μg/m^{3}和35 μg/m^{3}的标准,中国超过一半(51.95%)的城市空气质量不达标,而且$PM_{2.5}$浓度呈现明显的“冬高夏低”分布模式(图4(a))。研究发现[11,38],原因主要是冬季土壤干燥,地表植被覆盖少,地面扬尘容易进入空气中,且冬季中国北部大范围地区供暖燃烧产生大量污染性气体[39],因此冬季$PM_{2.5}$浓度最高,达到了71.02 μg/m^3。夏季降雨量最大,天气系统变化较强,$PM_{2.5}$浓度最低[40],为31.02 μg/m^3。春秋季由于天气系统转换,常伴随着不稳定的天气系统变化,气候扩散条件较好[41],$PM_{2.5}$浓度分别为40.42 μg/m^{3}和41.79 μg/m^3。

(2)月度$PM_{2.5}$分布

中国的$PM_{2.5}$浓度月度变化曲线呈“U”形(图4(b)),2−5月$PM_{2.5}$浓度呈快速下降趋势,6−9月$PM_{2.5}$浓度维持在较平稳水平,7月天气炎热,扬尘严重,因而$PM_{2.5}$浓度略微上升。10−12月$PM_{2.5}$呈现明显的上升趋势。1月$PM_{2.5}$浓度最高,达到77.31 μg/m^3,超过75 μg/m^{3}的浓度标准。2月、3月和10−12月的$PM_{2.5}$浓度在35 μg/m^{3}和75 μg/m^{3}之间。5月$PM_{2.5}$浓度下降为34.14 μg/m^3,8月浓度最低,为29.93 μg/m^3,5−9月$PM_{2.5}$浓度均在35 μg/m^{3}以下。

(3)逐小时$PM_{2.5}$分布

中国的$PM_{2.5}$逐小时浓度呈双峰变化(图5),双峰分布在上午10−12时和夜间21−22时,$PM_{2.5}$浓度分别达到了45.78 μg/m^{3}和44.86 μg/m^3。研究发现[39],10时开始地表太阳辐射增强,人群活动频率逐渐增加,污染排放开始积累,从而导致$PM_{2.5}$含量升高。另外,20时下班晚高峰和人群夜生活的影响致使夜间20−22时$PM_{2.5}$含量达到峰值。而下午由于温度升高,局部地表差异较大,空气对流增强,使得颗粒物浓度有所降低,最低为34.73 μg/m^3。3−6时,人类活动对空气质量的影响最弱,$PM_{2.5}$含量稳定并轻微降低。

4.1.2　空间分布分析

(1)中国$PM_{2.5}$空间分布分析

从本文基于一维线性Kalman的$PM_{2.5}$时空分布分析模型的的实验结果来看,中国$PM_{2.5}$的分布呈现明显的空间异质性[42]。从$PM_{2.5}$浓度年均值来看,中国$PM_{2.5}$浓度分布呈现以“新疆−华北平原”为中心的双核分布特征,这与王振波的结论[11]基本符合。如图6所示,作为双核分布的核心,新疆喀什地区以及

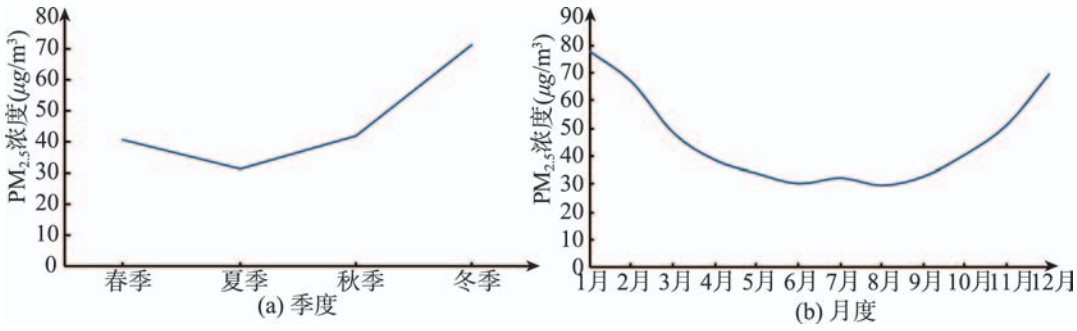

图4　中国$PM_{2.5}$季度和月度平均浓度

Fig.4　Quarterly and monthly average concentrations of $PM_{2.5}$ in China

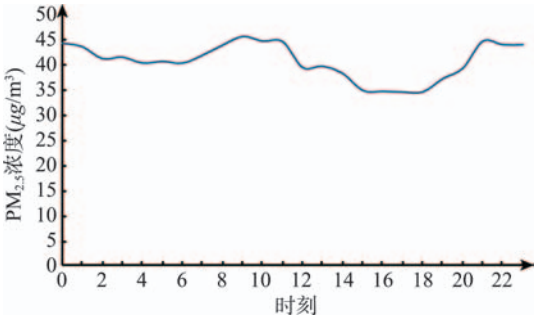

图5 中国$PM_{2.5}$逐小时浓度

Fig.5 Hourly concentration of $PM_{2.5}$ in China

华北平原的河北南部、山东西部、河南北部和山西东南部$PM_{2.5}$浓度高达70~80 μg/m³，并向四周扩散衰减。研究表明，河北、河南、山东等区域的常年$PM_{2.5}$高浓度分布主要来源于人为污染，重工企业高源排放产生大量大气污染物[43]，加之地形和气象要素的影响[39]，形成相互输送的重污染区[44]。核中心附近的新疆大部、辽宁、湖北、山西、安徽、江苏$PM_{2.5}$浓度达到了50~60 μg/m³，吉林、陕西、宁夏、四川、重庆、湖南、浙江等地$PM_{2.5}$年均浓度为40~50 μg/m³，青海、甘肃北部、内蒙古北部、黑龙江北部及广西、广东、江西等地区$PM_{2.5}$年均浓度维持在30~40 μg/m³，西藏、云南、海南、珠江三角洲地区、福建、空气质量良好，年均浓度为20~30 μg/m³，低于GB 3095-2012年均限值35 μg/m³。西藏、云南人口稀疏，开发强度不大，且植被覆盖度高，因此$PM_{2.5}$含量很低，空气质量高。海南、福建为沿海省份，由于空气强对流和海水的吸收作用，可吸入颗粒物含量极低。

同时，中国的$PM_{2.5}$浓度空间分布的分界线与以“黑河–腾冲”为界的胡焕庸线吻合度极高。在胡焕庸线东南的地区集中了中国绝大部分人口，同时，该地区的$PM_{2.5}$浓度也比胡焕庸线西北的青海、西藏等地区高。

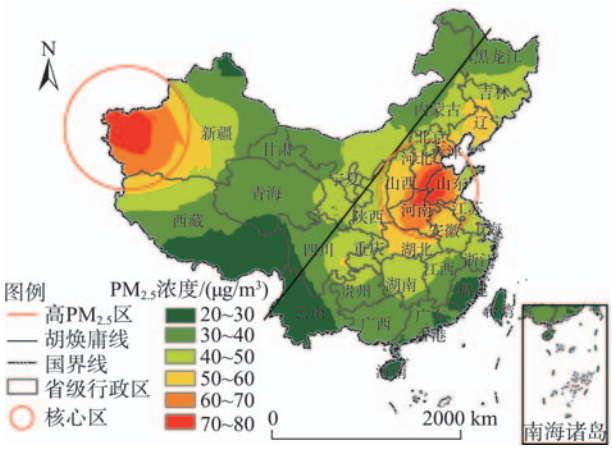

图6 2015年中国$PM_{2.5}$平均浓度空间分布图

Fig.6 Spatial distribution of average $PM_{2.5}$ concentration in China in 2015

（2）中国$PM_{2.5}$月度空间分布分析

图7表现了中国$PM_{2.5}$空间分布的月度变化特征。1月，中国超过50%的地区$PM_{2.5}$浓度超过75 μg/m³，尤其是华北平原和湖北，其月均$PM_{2.5}$浓度超过了100 μg/m³，部分城市超过120 μg/m³。2月全国$PM_{2.5}$浓度有所下降，但以湖北和华北平原为中心的地区空气中颗粒物月均浓度依旧在75 μg/m³以上。3–9月中国大部分地区空气质量较良好，而从4月开始新疆西北部的喀什地区出现较严重的颗粒物污染现象，且其浓度超过了125 μg/m³。薛江丽等[44]研究发现新疆在春季沙尘暴期间（3–5月）$PM_{2.5}$浓度明显上升，而且春季天气交替造成沙尘天气频繁，因此春季新疆$PM_{2.5}$浓度较高。10月开始，华北平原和新疆地区空气状况开始变差。11月，以吉林、辽宁为中心的东北地区空气中颗粒物含量急剧上升，超过了100 μg/m³。12月，重度污染区开始南移，甘肃以东、长江以北的地区几乎都受重度污染影响，而西北的新疆地区也维持较高的$PM_{2.5}$浓度。

4.2 K-Medoids聚类结果

本研究采用Kriging模拟出中国$PM_{2.5}$的月度空间分布差异，并通过地统计的方法获取每个城市的$PM_{2.5}$月度浓度。基于12个月的时序数据，本研究在地市级尺度上采用K-Medoids聚类方法对中国各城市进行$PM_{2.5}$浓度的时序聚类，来识别$PM_{2.5}$浓度变化具有相似时间分布特征的城市群。

本研究首先采用Silhouette指标来最优化聚类数目，在类间差异最大的情况下使类内数据具有最高的相似度。基于Silhouette指标的聚类数目选择结果如图8所示，在聚类数目为4时，Silhouette值达到峰值0.3256，表明将城市群聚集为4类能最大程度上划分其时间分布差异。因此，本研究选取中国各城市12个月的$PM_{2.5}$平均浓度作为特征，采用AP初始化聚类中心，并以DTW路径长度为特征距离，将城市群划分为4类。

图9反映了采用K-Medoids进行城市聚类后，4个类别的聚类中心。从图9可以明显地看出，第1类城市的月度$PM_{2.5}$浓度变化最大，且平均浓度最高，污染最严重，$PM_{2.5}$浓度除了在3月和5月存在小型的上升趋势，其一年的变化趋势基本符合“U”型

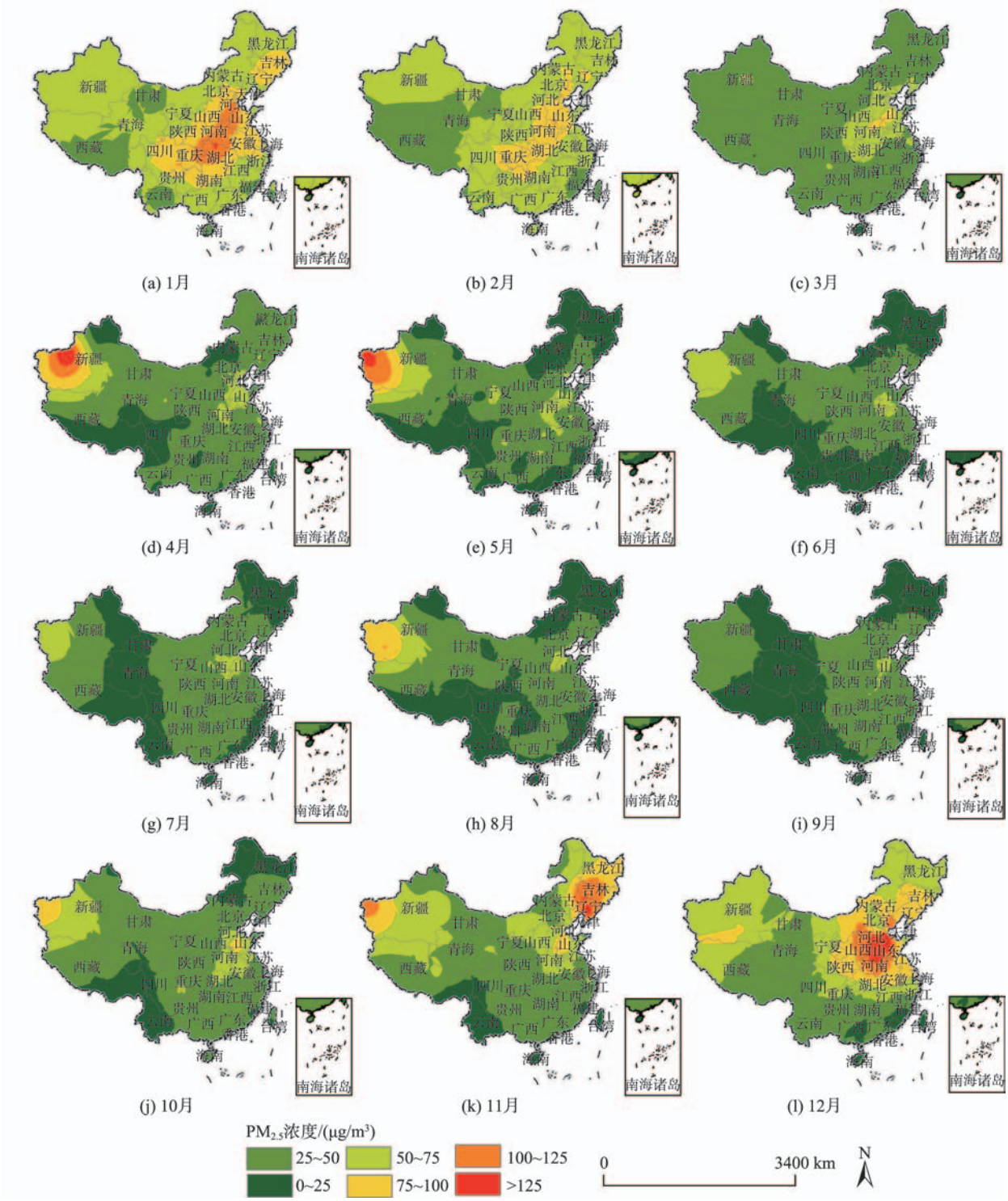

图7 中国2015年1–12月$PM_{2.5}$浓度空间分布图

Fig.7 Spatial distribution of $PM_{2.5}$ concentration in China in 2015

分布。第2类城市的$PM_{2.5}$平均浓度低于第1类,且月度变化不大,均处于15~50 μg/m³之间。第3类和第4类分布规律近似,且平均浓度最低,但第4类在1–8月浓度均低于第3类,而进入10月之后$PM_{2.5}$浓度反而高于第3类。

为了检验采用12个月$PM_{2.5}$浓度值作为特征的空间聚类结果,本研究采用q统计量[45–46]来评估$PM_{2.5}$的空间分异性。q统计量的计算公式为:

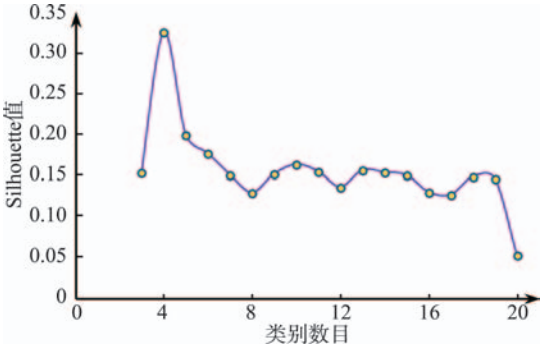

图8 Silhouette值随类别数目变化图

Fig.8 The variation of Silhouette values with the changes in number of categories

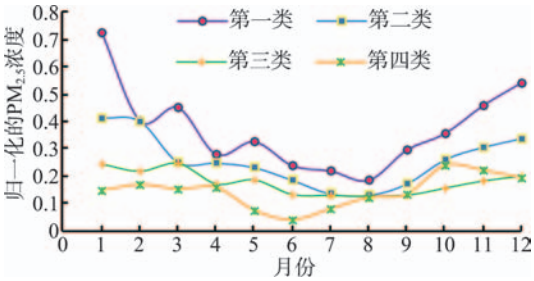

图9 聚类中心特征分布图

Fig.9 Features distribution of clustering centers

$$q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2} = 1 - \frac{SSW}{SST} \qquad (2)$$

式中：N为研究单元数目，此处为研究区内城市数目；h为分层数，此处为聚类类别数；SSW为层内方差，SST为总方差。结果表明(图10)，1−12月q统计量呈现先减小后增大的趋势，其中8月q统计值达到最低值0.137，而1月达到最大值0.645，这说明夏季尤其是8月$PM_{2.5}$的空间分异性不明显，而秋冬季$PM_{2.5}$的空间分异特性非常显著。

从图11可看出，$PM_{2.5}$的分布具有极大的集聚性，相同类别的城市分布具有较大的空间相关性。

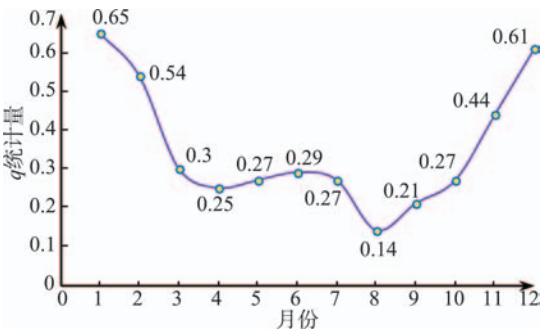

图10 q统计量分布图

Fig.10 q statistic distribution

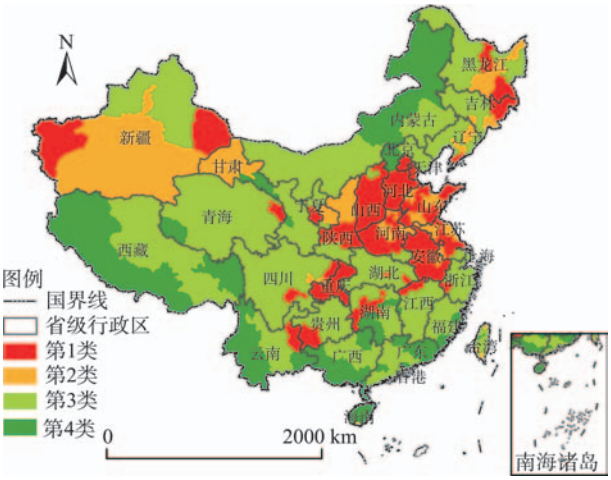

图11 基于K-Medoids的城市$PM_{2.5}$浓度聚类结果图

Fig.11 Clustering results of $PM_{2.5}$ concentrations of cities based on K-Medoids

结合图7和图9可知，中国绝大部分城市$PM_{2.5}$超标严重，热点区域(第1类)主要集中在华北平原、江淮、重庆、贵州南部、新疆西部和东北，该地区城市的$PM_{2.5}$浓度最高，污染最严重，且月度$PM_{2.5}$浓度变化大，浮动剧烈；空气质量相对最好的地区(第4类)主要分布在内蒙古北部、西藏南部、云南、广西、广东、福建，这些地区要么城市化程度不高，要么地理气象条件利于颗粒物扩散和吸收，因此$PM_{2.5}$浓度最低。此外，$PM_{2.5}$浓度次级高的地区(第2类)主要包括新疆大部、甘肃北部、陕西北部，以及华北平原重污染地区外围。前者虽然不是重工企业密集区，但植被覆盖度低，沙尘天气多，降雨量少，颗粒物扩散困难，因此$PM_{2.5}$浓度较高；而后者作为重污染区的缓冲地带，在大气系统的扩散和交换作用下，该地区空气质量受重污染区污染物的影响，$PM_{2.5}$浓度常年较高。第3类地区占据中国大半部分区域，$PM_{2.5}$浓度较良好，但在冬春季仍存在较频繁的污染情况。

5 结论

本研究利用网络大数据深度发掘出了$PM_{2.5}$的时空变化规律在数据量大、数据繁杂的情况下，用Kalman滤波对观测数据进行最佳估计和降维清洗，并利用空间Kriging插值和地统计方法模拟中国的$PM_{2.5}$时空分布，探讨其时空变异规律。此外，本研究基于DTW的K-Medoids聚类分析，探讨了中国各类城市$PM_{2.5}$的时空分布规律。

从时间维度分析，中国$PM_{2.5}$浓度呈现出春夏

低、秋冬高的变化模式,冬季远远高于夏季,部分地区浓度超过200 μg/m^3;日均$PM_{2.5}$浓度呈现以10−12时和21−22时为峰值的“W”形分布。从空间维度分析,中国超过半数地区的$PM_{2.5}$浓度超过国家标准(年度平均浓度不超过35 μg/m^3),且严重超标的地区主要分布在以华北、江淮平原和塔里木盆地为核心的地区,$PM_{2.5}$浓度以这些核心区域为中心向四周减弱分布,青藏高原、云贵、广西、广东、福建以及内蒙古东北部等地区的空气质量相对良好。

基于中国城市月度$PM_{2.5}$浓度的K-Medoids聚类结果表明,中国城市的$PM_{2.5}$时序分布具有较大的空间相关性,地理环境相似、经济发展产业结构相似的地区$PM_{2.5}$浓度具有较相似的时间分布特征,研究表明[47−48],$PM_{2.5}$的空间分布与地域、气象等因素密不可分,从城市尺度研究发现$PM_{2.5}$的分布与工商业发达程度、人口聚集程度有较大相关性[49]。聚类结果还显示重污染区主要集中在华北平原、江淮以及东北、新疆部分地区,空气质量最佳区主要分布在东南沿海和西南、内蒙古部分地区。

在今后的研究中,将耦合卫星遥感数据进行$PM_{2.5}$的地面观测值纠正,同时通过积累更多的数据进行不同尺度的研究,结合气象、地貌、社会经济结构等要素分析重污染区$PM_{2.5}$的成因,以提出科学的治理对策。

参考文献(References):

[1] 张智,白穆,游浩妍.基于MODIS数据的$PM_{2.5}$反演在大气污染监测中的应用[J].测绘科学,2016(9):1-10. [Zhang Z, Bai M, You H Y. Application of high spatial resolution $PM_{2.5}$ retrieval in air pollution monitor[J]. Science of Surveying and Mapping, 2016,9:1-10.]

[2] Pope Iii C A, Burnett R T, Thun M J, et al. Lung cancer, cardiopulmonary mortality, and long- term exposure to fine particulate air pollution[J]. Jama, 2002,287(9):1132-1141.

[3] Kappos A D, Bruckmann P, Eikmann T, et al. Health effects of particles in ambient air[J]. International Journal of Hygiene and Environmental Health. 2004,207(4):399-407.

[4] Mohammed M O, Song W W, Li W, et al. Potential toxicological and cardiopulmonary effects of $PM_{2.5}$ exposure and related mortality: Findings of recent studies published during 2003-2013[J]. Biomedical and Environmental Sciences, 2016,29(1):66-79.

[5] Liu Y, Paciorek C J, Koutrakis P. Estimating regional spatial and temporal variability of $PM_{2.5}$ concentrations using satellite data, meteorology, and land use information[J]. Environmental health perspectives, 2009,117(6):886.

[6] Lee H J, Liu Y, Coull B A, et al. A novel calibration approach of MODIS AOD data to predict $PM_{2.5}$ concentrations[J]. Atmos. Chem. Phys, 2011,11(15):7991-8002.

[7] Van Donkelaar A, Martin R V, Park R J. Estimating ground-level $PM_{2.5}$ using aerosol optical depth determined from satellite remote sensing[J]. Journal of Geophysical Research: Atmospheres, 2006,111(D21):5049-5066.

[8] Ma Z, Hu X, Sayer A M, et al. Satellite-based spatiotemporal trends in $PM_{2.5}$ concentrations: China, 2004- 2013[J]. Environ. Health Perspect, 2015,124:184-192.

[9] Paciorek C J, Liu Y. Limitations of Remotelysensed Aerosol as a Spatial Proxy for Fine Particulate Matter[R]. Harvard University Biostatistics Working Paper Series, Working Paper 89, 2008.

[10] 武装,覃爱明.基于大数据的空气质量数据可视化[J].中外企业家,2015(3):249:253. [Wu Z, Qin A M. Air quality data visualization based on big data[J]. Chinese & Foreign Entrepreneurs, 2015,3:249-253.]

[11] 王振波,方创琳,许光,等. 2014年中国城市PM_($_{2.5}$)浓度的时空变化规律[J].地理学报,2015,70(11):1720-1734. [Wang Z B, Fang C L, Xu G, et al. Spatial-temporal characteristics of the $PM_{2.5}$ in China in 2014[J]. Acta Geographica Sinica, 2015,70(11):1720-1734.]

[12] 潘红玲.中国重度雾霾时空分布特征及影响因子分析[D].成都:电子科技大学,2015. [Pan H L. Time and space distribution characteristics of the severe fog and haze of China and the influence factor analysis[D]. Chengdu: University of Electronic Science and Technology of China, 2015.]

[13] 李健军,杜丽,王晓彦,等. $PM_{2.5}$自动监测仪器第一阶段测试报告和技术指标要求[R].中国环境监测总站,2012. [Li J J, Du L, Wang X Y, et al. $PM_{2.5}$ automatic monitoring instrument in the first stage test report and technical index requirements[R]. China National Environmental Monitoring Station, 2012.]

[14] 宁爱民,文军浩,郑德智,等. $PM_{2.5}$监测技术及其比对测试研究进展[J].计测技术,2013(4):11-14. [Ning A M, Wen J H, Zheng D Z, et al. Advances in monitoring technologies and its comparison research for $PM_{2.5}$[J]. Metrology & Measurement Technology, 2013,4:11-14.]

[15] Kalman R E. A new approach to linear filtering and prediction problems[J]. Journal of basic Engineering, 1960, 82(1):35-45.

[16] Kalman R E, Bucy R S. New results in linear filtering and prediction theory[J]. Journal of basic engineering, 1961,

83(1):95-108.

[17] 李慧茹.基于kalman滤波的近实时电离层TEC监测与反演[D].西安:长安大学,2013. [Li H R. Near real-time monitoring and inverting TEC of ionosphere based on kalman filter[D]. Xi'an: Chang'an University, 2013.]

[18] 邱凤云. Kalman滤波理论及其在通信与信号处理中的应用[D].济南:山东大学,2008. [Qiu F Y. Kalman filtering with its application to communication and signal processing[D]. Jinan: Shandong University, 2008.]

[19] Kaufman L, Rousseeuw P. Clustering by means of medoids[M]. North-Holland, 1987.

[20] Kaufman L, Rousseeuw P J. Finding groups in data: An introduction to cluster analysis[M]. John Wiley \& Sons, 2009.

[21] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases[M]. Springer, 1993.

[22] Chan K, Fu A W. Efficient time series matching by wavelets[Z]. IEEE, 1999126-133.

[23] Megalooikonomou V, Wang Q, Li G, et al. A multiresolution symbolic representation of time series[Z]. IEEE, 2005:668-679.

[24] Perng C, Wang H, Zhang S R, et al. Landmarks: A new model for similarity-based pattern querying in time series databases[Z]. IEEE, 2000:33-42.

[25] Keogh E. A fast and robust method for pattern matching in time series databases[J]. Proceedings of WUSS, 1997, 97(1):99.

[26] Fu T. A review on time series data mining[J]. Engineering Applications of Artificial Intelligence. 2011,24(1):164-181.

[27] Berndt D J, Clifford J. Using dynamic time warping to find patterns in Time Series[Z]. Seattle, WA, 1994359-370.

[28] Liao T W. Clustering of time series data-a survey[J]. Pattern recognition, 2005,38(11):1857-1874.

[29] 迟妍妍,张惠远.大气污染物扩散模式的应用研究综述[J].环境污染与防治,2007(5):376-381. [Chi Y Y, Zhang H Y. A review of the development and application of air pollutant dispersion models[J]. Environmental Pollution & Control, 2007(5):376-381.]

[30] Zhang X, Liu J, Du Y, et al. A novel clustering method on time series data[J]. Expert Systems with Applications. 2011,38(9):11891-11900.

[31] 刘贤梅,赵丹,郝爱民.基于优化的DTW算法的人体运动数据检索[J].模式识别与人工智能,2012(2):352-360. [Liu X M, Zhao D, Hao A M. Human motion data retrieval based on dynamic time warping optimization algorithm [J]. Pattern Recognition and Artificial Intelligence, 2012 (2):352-360.]

[32] Frey B J, Dueck D. Clustering by passing messages between data points[J]. science. 2007,315(5814):972-976.

[33] Dueck D. Affinity propagation: Clustering data by passing messages[D]. Citeseer, 2009.

[34] 杨传慧,吉根林,章志刚. AP算法在图像聚类中的应用研究[J].计算机与数字工程,2012(10):119-121. [Yang C H, Ji G L, Zhang Z G. Research on application of algorithm AP in images clustering[J]. Computer and Digital Engineering, 2012,10:119-121.]

[35] Rousseeuw P J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of computational and applied mathematics. 1987,20:53-65.

[36] de Amorim R C, Hennig C. Recovering the number of clusters in data sets with noise features using feature rescaling factors[J]. Information Sciences, 2015,324:126-145.

[37] Llet I R, Ortiz M C, Sarabia L A, et al. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes[J]. Analytica Chimica Acta, 2004,515(1):87-100.

[38] 赵晨曦,王云琦,王玉杰,等.北京地区冬春$PM_{2.5}$和PM_{10}污染水平时空分布及其与气象条件的关系[J].环境科学,2014(2):418-427. [Zhao C X, Wang Y Q, Wang Y J, et al. Temporal and spatial distribution of $PM_{2.5}$ and PM_{10} pollution status and the correlation of particulate matters and meteorological factors during winter and spring in Beijing[J]. Environmental Science, 2014(2):418-427.]

[39] 李珊珊,程念亮,张玉洁,等. 2014年华北地区$PM_{2.5}$数值模拟研究: 2015年中国环境科学学会学术年会[Z].中国广东深圳:2015:7. [Li S S, Chen N L, Zhang Y J, et al. Numerical simulation research of $PM_{2.5}$ in north China in 2014[Z]. China Environmental Science Society Annual Conference Proceedings, 2015:7.]

[40] 关月,何立富. 2013年1月大气环流和天气分析[J].气象,2013,39(4):531-536. [Guan Y, He L F. Analysis of January 2013 atmosphere circulation and weather[J]. Meteorological Monthly, 2013,39(4):531-536.]

[41] Yang F, Tan J, Zhao Q, et al. Characteristics of $PM_{2.5}$ speciation in representative megacities and across China[J]. Atmospheric Chemistry and Physics, 2011,11(11):5207-5219.

[42] 郑玫,张延君,闫才青,等.中国$PM_{2.5}$来源解析方法综述[J].北京大学学报(自然科学版),2014(6):1141-1154. [Zheng M, Zhang Y J, Yan C Q, et al. Review of $PM_{2.5}$ source apportionment methods in China[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014(6):1141-1154.]

[43] 李珊珊,程念亮,徐峻,等. 2014年京津冀地区$PM_{2.5}$浓度时空分布及来源模拟[J].中国环境科学,2015(10):2908-

2916. [Li S S, Cheng N L, Xu J, et al. Spatial and temporal distrubions and source simulation of $PM_{2.5}$ in Beijing-Tianjin-Hebei region in 2014[J]. Chinese Environmental Science, 2015(10):2908-2916.]

[44] 薛江丽,李俊,张鑫,等.新疆春季两次沙尘暴过程中大气$PM_{2.5}$元素组成特征分析[J].环境与健康杂志,2010(9):759-763. [Xue J L, Li J, Zhang X, et al. Characteristics of elemental compositions of ambient $PM_{2.5}$ during sandstorm in spring in Xinjiang[J]. Environ Health, 2010(9):759-763.]

[45] Wang J, Li X, Christakos G, et al. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China[J]. International Journal of Geographical Information Science, 2010,24(1):107-127.

[46] Wang J, Zhang T, Fu B. A measure of spatial stratified heterogeneity[J]. Ecological Indicators,2016,67:250-256.

[47] 张佟佟,李茜,张建辉,等. $PM_{2.5}$污染特征研究综述:2014中国环境科学学会学术年会[Z].中国四川成都:2014:5. [Zhang T T, Li Q, Zhang J H, et al. A review on $PM_{2.5}$ pollution characteristic research[Z]. China Environmental Science Society Annual Conference Proceedings, 2014:5.]

[48] 付桂琴,张迎新,谷永利,等.河北省霾日变化及成因[J].气象与环境学报,2014,30(1):51-56. [Fu G Q, Zhang Y X, Gu Y L, et al. Change of haze day and its forming reason in Hebei province[J].Journal of Meteorology and Environment, 2014,30(1):51-56.]

[49] 郭涛,马永亮,贺克斌.区域大气环境中$PM_{2.5}$/PM_{10}空间分布研究[J].环境工程学报,2009(1):147-150. [Guo T, Ma Y L, He K B. Study on spatial distributions of $PM_{2.5}$/PM_{10} in regional atmospheric environment[J]. Chinese Journal of Environmental Engineering, 2009,1:147-150.]